MOLECULAR ECOLOGY

Molecular Ecology (2010), **19** (Suppl. 1), 147–161 doi: 10.1111/j.1365-294X.2009.04484.x

Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack

CHARLES H. CANNON,*† CHAI-SHIAN KUA,* D. ZHANG* and J.R. HARTING†

Ecological Evolution Group, Xishuangbanna Tropical Botanic Garden, Chinese Academy of Sciences, Menglun, Mengla 666303, China, †Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA

Abstract

Most comparative genomic analyses of short-read sequence (SRS) data rely upon the prior assembly of a reference sequence. Here, we present an assembly free analysis of SRS data that discovers sequence variants among focal genomes by tabulating the presence and frequency of 'complex' fragments in the data. Using data from nine tree species, we compare genomic diversity from populations to families. As a control, we simulated SRS data for three known plant genomes. The results provide insight into the quality and distributional bias of the sequencing reaction. Three main types of informative complexmers were identified, each possessing unique statistical properties. Type I complexmers are unique to a genome but suffer from a high false positive rate, being highly dependent on read coverage and distribution. Type II complexmers are shared between two genomes and can highlight potential copy-number differences. Type III complexmers are exclusive to a subset of genomes and can be useful for associating genetic differences with phenotypic or geographic variation. At the population level in an endangered timber species, numerous markers were identified that could potentially determine geographic origin of individuals and regulate international trade. We observed that the genomic data for the four fig species were more divergent than for stone oak species, possibly due to their complex pollination syndrome and high rates of gene flow. Our approach greatly enhances the application of SRS technology to the study of non-model organisms and directly identifies the most informative genetic elements for more detailed study and assembly.

Keywords: complexmers, simulated SRS data, stone oaks, *Lithocarpus*, figs, *Ficus*, ramin, *Gonystylus*

Received 15 June 2009; revision received 28 September 2009; accepted 13 October 2009

Introduction

Since the introduction of short-read genome sequencing technologies, analytical techniques have focused entirely on assembly based approaches (Pop & Salzberg 2008), where the raw data is either aligned against a closely related reference genome (Li *et al.* 2008; Langmead *et al.* 2009) or against *de novo* contigs and scaffolds assembled using a variety of algorithms (Warren *et al.* 2007; Zerbino & Birney 2008; Simpson *et al.* 2009). van Bers *et al.* (2009) effectively discovered a large number of potentially informative SNPs by using de novo assembly approaches and then mapping these against a reference genome. Here, we present a comparative genomics approach for short-read sequence (SRS) data based upon a direct assembly free survey of the 'complex' fragments (complexmers) within the sequence data. Our approach requires no prior knowledge of the focal genomes and allows comparisons between any two organisms. We illustrate the flexibility of our approach by examining whole genomic SRS data from nine species of tree, mostly tropical Asian taxa, spanning population to family level comparisons. We simulated SRS data from the physical maps of three finished plant genomes to serve as controls for our analysis and to provide reference to our interpretation of the resulting patterns.

The genomes in our study include five species of Fagaceae (primarily of the stone oaks, *Lithocarpus*), four species of fig (Moraceae:*Ficus*), and an endangered spe-

Correspondence: Charles H. Cannon, Fax: +861698715070, E-mail: chuck@xbtg.ac.cn

cies of tropical timber, commonly called 'ramin' (Thymelaeaceae: *Gonystylus bancanus*), represented by five individuals each from a different population. These genomes form three projects, aimed at different fundamental questions about speciation, diversification, and regulation of international markets (see Methods). The species in these three projects are all ecologically and/or economically important in the tropical rainforests of Southeast Asia and yet virtually nothing is known about their genomic biology.

Although the power and accuracy of *de novo* assembly algorithms continues to improve, these approaches are much more effective for haploid genomes (Cronn *et al.* 2008; Reinhardt *et al.* 2009) and have difficulty with duplicated portions of the genome (Phillippy *et al.* 2008) or highly heterozygous genomes. Our unpublished results indicate that a large portion of the DNA sequence data from our diploid and probably highly out-crossed tree species does not assemble, despite efforts to optimize the parameters of the reconstruction. Additionally, the assembled *de novo* contigs are strongly biased towards conserved regions with low levels of polymorphism. These limitations are fundamentally difficult problems for studying poorly known organisms, especially when high levels of heterozygosity, historical polyploidization and segmental duplication are major features of the species biology. Given the enormous effort required to 'complete' a genome, the availability of reference genomes will be a chronic problem for ecologists.

To address the many compelling ecological and evolutionary questions among diverse but poorly known tropical organisms using genomic techniques, novel approaches for analysing SRS data are obviously necessary. In the tropics, many of the more compelling questions are related to the evolution and maintenance of biodiversity. To incorporate these issues into genomic studies, a comparative approach is required (Ellegren 2008), using numerous exemplar species from a 'model group' that exhibits important natural variation. With the promise of rapidly increasing output from the DNA sequencing technology (Pushkarev *et al.* 2009), techniques for rapidly and efficiently comparing dozens of closely related species are needed. Our assembly free approach discovers the most informative genetic elements that both distinguish and unite any subset of genomes, prior to assembly of the SRS data. These genomic elements could be used directly as a DNA fingerprinting device, a highly sensitive multi-locus DNA barcode, or in association tests for quantitative traits. Additionally, *de novo* contigs can be assembled in a highly targeted and local fashion and these short fragments used as reference sequences for alignment. Basically, our approach performs an *in silico* DNA–DNA hybridization analysis of only the complex fraction of the genome. We refer to these complex sequence fragments as 'complexmers' Similar to a genome subtraction experiment or representational difference analysis (Lisitsyn *et al.* 1993), the results are not limited to just the differences between genomes but also the similarities.

Here, we address several basic questions, both in terms of the analytical approach and the broad biological and ecological questions framed by our focal genomes. What are the basic properties of complexmers in both real and simulated SRS data? What magnitude and scope of genomic differences are revealed? What types of patterns are revealed from pair-wise correlation of complexmer frequency? How well do the evolutionary relationships identified among genomes by this approach match known relationships? Can the ecologically and evolutionary informative complexmers be assembled into larger fragments? Can the genomic differences discovered be interpreted in relation to the important ecological and evolutionary traits of the study organisms?

Materials and methods

Study groups

Fagaceae. Trees in this family are ecologically prominent in the tropical forests of Asia (Young & Herwitz 1995; Corlett 2007; Hua 2008), with two ecologically dominant and taxonomically diverse genera (*Lithocarpus* and *Castanopsis*). The current distribution of these trees (Soepadmo 1972; Huang *et al.* 2000) indicate that they are sensitive to seasonal rainfall and temperature, as they do not extend beyond the subtropical zone in China and they do not persist in the seasonally dry Lesser Sunda Islands of Indonesia. They are often used in the pollen record to indicate cooler wet climates in the past (Morley 2000; Zheng & Li 2000; Anshari *et al.* 2001). For these reasons, the family is a good model group for the historical distribution of evergreen tropical forests in Southeast Asia (Cannon & Manos 2003) and the future response to global climate change. The biogeographic history of Southeast Asian forests is particularly important (Petit *et al.* 2008) because their current distribution is unrepresentative of its past (Cannon *et al.* 2009). Additionally, the evolution of fruit morphology (Cannon & Manos 2000, 2001) is under strong selection pressure from rodent seed predation and dispersal (Xiao *et al.* 2005, 2006). Our samples included four stone oaks (*Lithocarpus*) and two outgroups at varying phylogenetic distances from this focal genus (Table 1). The data for this study have been archived at the NCBI Short Read Archive under the accession number SRA00938.2.

Ficus. The Moraceae includes 35 genera but two genera (*Ficus* and *Artocarpus*) are the most ecologically predom-

Table 1 Amount and type of short-read sequence data, including the total number of unique and frequent 25 base pair complex-mers. Location: 1) Sumatra, Indonesia; 2) Terengganu, Malaysia; 3) Johor, Malaysia; 4) Pahang, Malaysia; 5) California, USA; 6) Sabah, Malaysia; 7) Yunnan, China; 8) XTBG, China. Type: S = Single-end library; P = Paired-end library

	Loc	Type	Total reads	Total bp	Complex-mers
Thymelaeaceae					
Gonystylus bancanus	1	S	18,265,251	507,426,938	1,116,663
Gonystylus bancanus	2	S	17,694,737	566,762,426	2,770,704
Gonystylus bancanus	3	S	20,126,067	646,046,751	2,665,279
Gonystylus bancanus	3	P	76,954,974	3,864,280,289	24,097,852
Gonystylus bancanus	4	S	12,487,609	353,399,335	737,232
Fagaceae					
Chrysolepis chrysophylla	5	S	13,876,056	396,855,202	1,268,288
Lithocarpus havilandii	6	S	20,760,407	575,063,274	301,510
Lithocarpus turbinatus	6	S	21,711,713	673,063,103	2,193,637
Lithocarpus hancei	7	P	154,235,816	7,293,259,431	19,589,907
Lithocarpus xylocarpus	7	P	66,494,720	3,356,506,909	27,125,479
Trigonobalanus diochangensis	7	P	153,785,918	7,356,935,517	27,734,580
Moraceae					
Ficus altissima	8	P	40,275,230	2,006,769,928	14,954,912
Ficus fistulosa	8	P	20,856,742	1,035,134,616	6,651,388
Ficus microcarpa	8	P	43,814,974	2,177,270,955	12,221,478
Ficus tinctoria	8	P	176,241,836	8,949,776,305	42,554,406
Simulated					
Arabidopsis thaliana		P	52,000,000	2,652,000,000	69,970,022
Oryza sativa		P	52,000,000	2,652,000,000	55,515,642
Populus trichocarpa		P	52,000,000	2,652,000,000	58,632,095

inant in tropical Asia. Figs (*Ficus*) have long been identified as 'keystone' species in tropical forest ecosystems (Janzen 1979; Harrison 2003, 2005), as they provide food to a wide range of frugivores throughout the year and help maintain animal populations in disturbed habitats. Figs also possess a complex co-evolutionary relationship with their pollinators (Ronsted *et al.* 2005), the highly specialized fig-pollinator wasps (Weiblen 2002). This symbiotic relationship has been the focus of a large amount of work (Machado *et al.* 2005; Herre *et al.* 2008; Jackson *et al.* 2008). Additionally, figs possess two main sexual systems (Weiblen 2000, 2004; Harrison & Yamamura 2003): 'monoecy' where unisexual flowers of both sexes are found on the same plant and 'dioecy' where unisexual flowers are found on different plants. These sexual systems should impose substantially different dynamics on population level gene flow, particularly in relation to effective population size. The exemplar species have been chosen to provide an initial comparison among growth forms and sexual systems (Zhou & Gilbert 2003). *Ficus microcarpa* and *F. altissima* (both subg. Urostigma) are large, monoecious hemi-epiphytes, while *F. tinctoria* (subg. Sycidium) and *F. fistulosa* (subg, Sycomorus) are small dioecious shrubs. *F. tinctoria* is highly variable in growth form, with one subspecies a true epiphyte (Zhou & Gilbert 2003). These four species are widespread, generally found throughout the Asian tropics and sub-tropics. More species are currently being sequenced. The data for this study have been archived at the NCBI Short Read Archive under the accession number SRA00939.1

Ramin. Gonystylus bancanus (Miq.) Kurz (family Thymelaeaceae) is limited to peat swamp forests along the inner margins of the South China Sea. The family Thymelaeaceae is a basal member of the Malvales, a large angiosperm order, which contains numerous economically valuable crops, including cotton and chocolate (Stevens 2001 onwards). The tropical peat swamps of Southeast Asia are widely recognized as containing a highly endemic and specialized biota (Rieley & Page 1995; Cannon & Leighton 2004), due to the extreme environmental conditions present. The over-exploitation and illegal smuggling of ramin has caused extensive damage to this fragile habitat and the species was placed on CITES Appendix II under the Convention on International Trade in Endangered Species, making it a major focus of international timber organizations (Wyn *et al.* 2004). We previously applied a novel DNA microarray-based technique to discover genomic signatures in different populations of this endangered species (Cannon *et al.* 2006) with the purpose of developing genetic markers for the determination of taxonomic identity and geographic origin of harvested timber.

Such markers would provide powerful tools for both producers and consumers of natural resources (Deguilloux *et al.* 2002; Nielsen & Kjaer 2008; Smulders *et al.* 2008). The data for this study has been archived at the NCBI Short Read Archive under the accession number SRA003731.1.

DNA samples and sequence data. Several methods were used to extract whole genomic DNA from the samples. Fresh leaf material was used for five Fagaceae, four *Ficus*, and the Sumatran individual of *G. bancanus*. DNA was extracted from 3–5 g of inner bark tissue from the four *G. bancanus* individuals from peninsular Malaysia. Two types of commercially available DNA extraction kits were used for most leaf samples: Qiagen Plant DNeasy Extraction Kits and MN NucleoSpin Plant DNA Extraction Kit. The *Chrysolepis chrysophylla* sample was isolated using a modified protocol for chloroplast-enriched extractions (Bookjans *et al.* 1984). Genomic DNA was extracted from *G. bancanus* using a modified C-TAB method (Doyle & Doyle 1987). The DNA samples were visualized and quantified on a check gel before shipment to the sequencing facility at the Micheal Smith Cancer Institute in British Columbia, Canada. Detailed sample information can be found in OSM 1.

At Canada's Micheal Smith Genome Sciences Centre sequencing facility, the genomic DNA samples were sonicated for 10 min and run on a 12% PAGE. The 400 bp DNA fraction was excised and eluted from the gel slice overnight at 4 °C in 300 μL of elution buffer [5:1, LoTE buffer (3 mM Tris–HCl, pH 7.5, 0.2 mM EDTA)-7.5 M ammonium acetate] and purified using a QIAquick purification kit (Qiagen, Cat#28104). Single-end (SET) and paired-end (PET) sequencing libraries (see Table 1) were constructed using Illumina genomic DNA prep kit by following company protocols (Illumina, cat# FC-102-1002). Clusters were generated on the Illumina cluster station (Illumina, cat# FC-103-1002) and sequence was run on Illumina 1G Analyzer following the manufacturer's instructions (Illumina, cat# FC-104-1003). The SET reactions generated two files, one containing the base calls and the other containing the quality of those base calls. R. Warren provided a script to trim the low quality data, using these two files. The PET reactions generated a single file for each member of the paired reads, using Illumina's GAPipeline-1.32, with the base calls and their associated quality scores. Sequence data was extracted from the PET quality files directly using an awk script.

To serve as a control, simulated Illumina data was generated from three known plant genomes: (thale cress:*Arabidopsis thaliana*; rice:*Oryza sativa*, and poplar:*Populus trichocarpa*). The most recent builds for these genomes, including the cytoplasmic genomes when available, were downloaded from the NCBI Genomic Biology website. Two data sets for each known genome containing 13 million paired-end 51 base pair reads were generated to simulate two typical lanes of Illumina PET data. Because preliminary results indicated that the cytoplasmic genomes were abundant in the Illumina data obtained from our samples, 20 copies of the cytoplasmic genomes were mixed with a single copy of the nuclear chromosomes during the simulation process.

Defining complexmers. All alignment and *de novo* assembly algorithms create an index or hash table (Stromberg & Marth 2007; Warren *et al.* 2007; Langmead *et al.* 2009; Simpson *et al.* 2009), using various *k*-mer fragment lengths, to efficiently organize or graph the sequence information in the data. These fragment lengths are always much shorter than the reads themselves. Ultimately, the choice of *k*-mer length is a trade-off between sensitivity and specificity. Short *k*-mers will provide a large amount of relatively unspecific information while long *k*-mers will provide a small amount of highly specific information. After a preliminary analysis and given the relatively broad phylogenetic scope of the current analysis, we chose to use 25 base pairs as our *k*-mer length. Ultimately, the length of complexmer should be optimized for each analysis.

Because highly repetitive DNA sequences are inherently less informative for comparative genomic purposes, we only analysed 'complex' *k*-mers. We used a simple definition for sequence complexity, based on three criteria: (1) no nucleotide can comprise more than half of the fragment; (2) all nucleotides must be present at least two times; and (3) no mono-nucleotide repeats longer than 4 base pairs were allowed. These complex 25 base pair fragments are subsequently referred to as 'complexmers'. Given 20 times coverage and a stochastic distribution of reads, a single nucleotide polymorphism would generate a different number of complexmers from each read (Fig. 1a). The mean number would be roughly equal to half the length of the complexmer (12.5 in this case) and a normal distribution of complexmer frequency centred on the SNP.

We will indicate the complex fraction of the genomic data for sample 'A' based upon this 25 base pair window by the symbol A^{cx25}. In a side-by-side evaluation of the three most common next-gen sequencing platforms, almost three-quarters of the sequencing error found in Illumina data was associated with repetitive sequence elements (Harismendy *et al.* 2009). Limiting the analysis to the complex fraction of the data will eliminate many of these sequencing errors. To eliminate single random sequencing errors, each complexmer had to be 'frequent' in the data, i.e. in at least three separate reads from a single lane, to be included in the analysis.

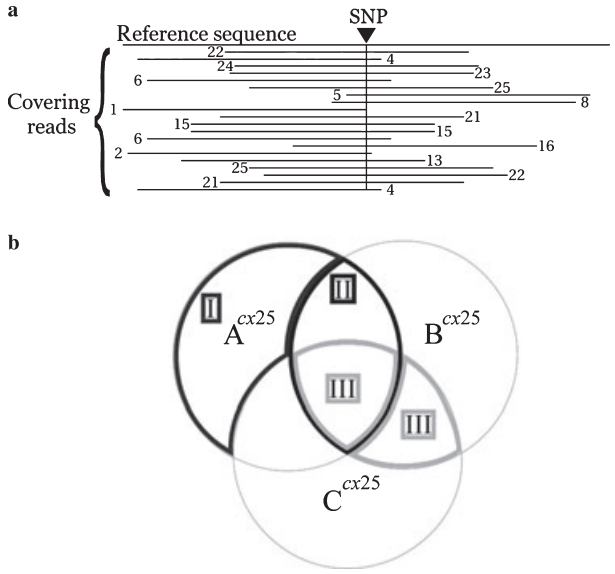

Fig. 1 Complexmers in short-read sequence data and in comparative genomic analysis. (a) The reference sequence is shown in black and the reads are shown in purple. Twenty 51 bp reads randomly cover a single nucleotide polymorphism, shown by the red triangle. The number of 25 bp complexmers in each read containing the SNP is shown at the end of each read. (b) Three types of complexmers with different statistical properties, given a genomic comparison among genomes A, B and C. The *cx*25 symbol indicates that the comparison includes only the complex fraction of the data, given a 25 base pair window. Type I. fraction unique to a single genome; II. fraction common to two genomes; and III. fraction shared exclusively by the genomes.

Currently, this lane by lane approach misses rare but valid sequence variants.

We tabulated the presence and frequency of all complexmers in each data file (a single lane on an Illumina flow cell), retaining those fragments observed three times or more (simple Python scripts and a more detailed description of the steps involved can be obtained from: www. ecologicalevolution.org/resources/). As unique complexmers were encountered, the reverse complement was counted simultaneously and the frequencies of these two equivalent sequences merged. This tabulation of complexmers was also performed for the simulated data sets. The tables from all of the data files for a sample were then merged together, accumulating the frequency of the complexmers across all data files.

Assembly free assessment of data quality. Quality assessment of the DNA sequencing reaction for species without a reference genome available is difficult. The frequency distribution of complexmer abundance provides insight into the overall quality of the genomic data, particularly in comparison to the results from simulated sequencing of known genomes. If the distribution of the reads is truly random, as it is in the simulated SRS data, the abundance of complexmers in the data would be proportional to their abundance in the genome. Given this ideal distribution of reads on the genome, few fragments will be super-abundant and a vast majority of the complexmers will be extremely rare as they only appear a single time in the genome. If, on the other hand, the distribution of the reads on the genome is strongly biased towards a small fraction of the genome, some complexmers will be super-abundant and rare complexmers will comprise a smaller fraction of the data. Each lane can be examined separately or all of the data from the individual can be combined in this analysis.

Assembly free comparative genomic analysis. Three types of informative complexmers can be directly identified in this comparative genomic framework, prior to assembly (Fig. 1b). Each type provides data of different quality and limitations.

Type I. Complexmers unique to a genome, defined as,

$$\{A^{Icx25} : A^{Icx25} \in A^{cx25}, A^{Icx25} \notin B^{cx25} \cap C^{cx25}\},$$

where A,B,C represent the available data for three different genomes, $x \in y$ means that x is a member of y, $x \notin y$ means that x is not a member of y, and $x \cap y$ means the union of x and y. Type I complexmers provide positive evidence for presence in one genome but inconclusive evidence for absence in other genomes, unless the complex fraction of all genomes are fully covered at least three times. Otherwise, it is possible that it exists in the other genomes but was simply not sequenced. To minimize false positives, we used an extremely high frequency threshold of fifty times (50 ×).

Type II. Complexmers shared by two or more genomes, defined as,

$$\{AB^{IIcx25} : A^{cx25} \cup B^{cx25}\},$$

where $x \cup y$ means the intersection of x and y. These type II complexmers provide positive evidence for presence in a pair of genomes. The correlation of complexmer frequency can be compared between genomes to understand the similarity in complexmer frequency in the two genomes. Additionally, a ratio test, similar to expression ratios in DNA microarray experiments (Amaratunga & Cabrera 2004), can used to identify complexmers which are significantly more abundant in one genome. Here, we standardize the distribution of ratios until the mean is zero and fit a normal distribution to determine frequency outliers in each genome.

Type III. Complexmers exclusive to two or more genomes, defined as,

$$\{ABC^{IIIcx25} : A^{cx25} \cup B^{cx25} \cup C^{cx25} \notin D^{cx25}\}$$

where D is an additional fourth genome while $x \cup y \cup z$ means the intersection of x,y and z $x \cup y \cup z \notin w$ means that the intersection of x,y and z is not a member of w. These type III complexmers provide positive evidence for conserved complexmers present in two or more genomes but absent in all other genomes. As with Type I complexmers, false positives because of poor coverage in some of the genomes will be an issue.

de novo assembly of informative polymorphisms. Because each well-covered polymorphism will generate many unique complexmers, depending on the depth and pattern of coverage (Fig. 1a), the *de novo* assembly of these complexmers is necessary to reconstruct the informative polymorphism. The successful assembly of complexmers will confirm two things: the polymorphism is adequately covered and the complexmers are homologous, i.e. coming from the same region, instead of representing small identical fragments with different flanking sequences. If the polymorphism is a single nucleotide and not close to any other polymorphisms, assembly of the complexmers should generate a DNA fragment equal to 2*(complexmer length in base pairs)-1. If the polymorphism involves an indel or a series of SNPs within the complexmer window, the assembled fragment could be longer. We performed this targeted *de novo* assembly technique, using ABySS (Simpson *et al.* 2009), on two sets of complexmers: for each sample, the Type II complexmers that were significantly more frequent in at least one pair-wise comparison were pooled; Type III complexmers exclusive to any subset of genomes. Non-default parameters for ABySS were *k*mer length (-*k* 13) and read length (-l 25). Only *de novo* contigs longer than 48 base pairs were retained. These were fused into a single pseudochromosome by separating them with a string of 35 N's and the combined reads for each sample were then aligned using *bowtie* (Langmead *et al.* 2009). Because the contigs are very short and alignment of 51 base pair reads against 50 base pair contigs is very poor, we adopted a strategy where three alignments were performed: one using the entire length of the reads and two where 15 base pairs were dropped from either end of the reads. This greatly improved the quality of alignments against the shortest contigs. While an exhaustive study of the resulting alignments of the sample data against these contigs discovered through our approach is not possible at this time, the highest quality alignments, given median coverage and average polymorphism, were tabulated to illustrate the power of these Type III complexmers in discovering differences and commonalities among the target genomes.

Results

Sequence data and complexmer diversity

For the fifteen samples of whole genomic sequence data, we obtained a total of 857 582 050 reads and 39 758 550 979 base pairs of DNA sequence data, after filtering out low-quality base calls (Table 1). The paired end (PET) reactions generated several fold more data per reaction than single end (SET) reactions. The six Fagaceae samples comprise the largest set of comparative data, with 430 864 630 reads and 19 651 683 436 base pairs, while the ramin samples, mostly consisting of SET reads, were the smallest with 145 528 638 reads and 5 937 915 739 base pairs. The single sample with the largest amount of available data is *Ficus tinctoria* with 176 241 836 reads and 8 949 776 305 base pairs, which should provide roughly 10 × coverage, given our estimate of its genome size.

Given the merged complexmer tabulations obtained for each sample, the number of frequent complexmers observed has a roughly asymptotic relationship with the total amount of sequence data (Table 1; Fig. 2a). Two of the Fagaceae samples have unusually low values for unique complexmers, given the amount of data, suggesting either the sequence data for these two samples is of lower quality than the other samples, the genome complexity for these taxa is substantially lower, or genome size substantially larger. The number of frequent complexmers observed in the simulated data from the three known genomes was nearly two times greater, given the amount of sequence data. The fact that *Arabidopsis thaliana* produced substantially more frequent complexmers is a result of the frequency threshold for each lane of data, i.e. each complexmer must be observed at least three times in a single lane of data. Because of the very small size of the *A. thaliana* genome, a single lane of data generated on the Illumina platform will simply provide higher overall coverage, thus a larger fraction of the overall complexmer population will be observed three times and be included among the frequent complexmers.

While this 'frequency' threshold prevents the constant introduction of random sequencing errors in the analysis, it also eliminates valid complexmers which are singletons within a single lane of Illumina data. We restricted the tabulation of complexmers to single lanes of data primarily because of current limitations of computer capacity, as the pooled complexmer table for the larger samples exceeds 10GB. A single simultaneous tabulation of complexmer frequency, using current tech-

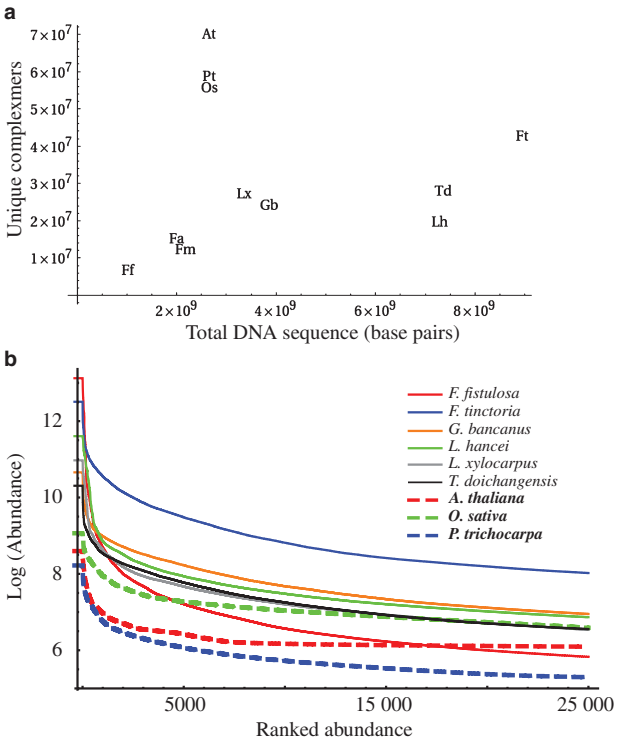

Fig. 2 Complexmer diversity in actual and simulated genomic data. (a) Relationship between total amount of DNA sequence available for a sample and the number of unique 25 bp complexmers present. Complexmers had to be present three times or more in at least one lane of data to be included in the analysis. Only samples with paired-end data are shown. Sample values are indicated by the following codes: At – *Arabidopsis thaliana*; Os – *Oryza sativa*; Pt – *Populus trichocarpa*; Gb – *Gonystylus bancanus*; Fa – *Ficus altissima*; Ff – *F. fistulosa*; Fm – *F. microcarpa*; Ft – *F. tinctoria*; Lh – *Lithocarpus hancei*; Lx – *L. xylocarpus*; Td – *Trigonobalanus doichangensis*. (b) Dominance diversity curves of complexmer abundance. The log abundance of the 25 000 most abundant complexmers are shown in ranked order, most abundant to the left. The value for the most abundant complexmer for each sample is shown by a horizontal extension over the *y*-axis.

niques, would require prohibitive amounts of RAM. Software to overcome these limitations is currently under development. As genome size increases, the fraction of singleton but valid complexmers increases, thus a larger fraction of valid complexmers would be eliminated in the current analysis.

We only have rough estimates for the genome size of our focal taxa but they range between 700 and 1000 Mbases, which are roughly twice the size of the largest known genome used in this analysis. Given the stark difference in the results between the simulated and actual SRS data, the genomic sequence data probably does not adequately cover the entire genome of our samples, despite the fact that the total amount of sequence data for several samples should provide roughly 10 × coverage. Despite this limitation, our analysis generated a very large population of 'frequent' complexmers for comparative analysis purposes. We simply need to be conservative in the identification of potential informative markers, particularly for Type I complexmers, which are prone to false positives.

The dominance diversity curves of the frequent complexmers also provide an indication of the overall quality of the sequence data (Fig. 2b). The most abundant complexmer in the three known genomes are orders of magnitude lower than in the actual data, despite the fact that these genomes are smaller and therefore have higher levels of sequence coverage. Interestingly, the relationship among the known genomes is not related to genome size, as the curve for rice is consistently higher than either thale cress or poplar. The curve for poplar, on the other hand, drops off quickly, indicating the high abundance complexmers are very rare in the simulated genomic data. For the SRS data collected in this study, the *Ficus fistulosa* sample has the most abundant complexmer but the curve drops quickly down so that it has the least abundant complexmers among all samples. This result probably indicates a bias in the overall distribution of reads to a relatively smaller portion of the overall genome. This type of bias has been observed before in SRS data (Dohm *et al.* 2008). The curve for *Ficus tinctoria*, on the other hand, starts high and remains high, which agrees with the large number of unique complexmers observed in the data. The patterns are less clear among the Fagaceae samples, as the very low number of unique complexmers observed for *Lithocarpus hancei* (Fig. 2a) was not reflected in an unusual dominance diversity curve. The curves for *L. xylocarpus* and *Trigonobalanus doichangensis* were almost identical, which also matches the total number of unique complexmers.

Type I complexmers

For many samples, an overwhelming fraction of the frequent complexmers observed in the genomic data were unique to an individual sample (Table 2). For example, over 92% of the complexmers observed in the *Trigonobalanus doichangensis* sample were unique, while almost 90% of those observed in the *Ficus tinctoria* sample were unique. Among the SET samples, by contrast, the Type I complexmers were much less frequent, always representing less than 50% and in many cases less than 10% of the total complexmers observed in the sample. Among congeneric samples with paired-end reads, the fraction of Type I complexmers was generally between 60% and 70%.

Table 2 Frequency of potential genetic markers among the different focal groups

	Type I complexmers		Type II *de novo* contigs		
	Total no.	50 × (%)	Total no.	Mean size (bp)	Max size (bp)
*Gb*1	114 681	3388 (3.0)	846	65	378
*Gb*2	189 295	5395 (2.9)	351	54	78
*Gb*3s	188 334	5273 (2.8)	430	56	124
*Gb*3p	20 633 987	56 677 (0.3)	1042	61	162
*Gb*4	4880	86 (1.8)	42	56	96
Cc	616 940	7790 (1.2)	1831	64	185
Lhav	31 790	425 (1.3)	434	58	155
Lt	585 769	7307 (1.2)	1858	59	366
Lhan	13 014 734	252 614 (1.9)	5919	62	452
Lx	20 169 409	137 743 (0.7)	4259	59	375
Td	25 557 643	701 435 (2.7)	3572	67	368
Fa	10 169 792	76 254 (0.7)	6177	71	412
Ff	4 372 975	131 806 (3.0)	1736	61	593
Fm	7 842 056	86 399 (1.1)	3575	62	316
Ft	38 229 911	1 169 895 (3.1)	6475	74	461

Type I complexmers: the total number of frequent complexmers observed in the genomic data for a single sample and the number of complexmers observed more than fifty times are shown in the left two columns (percentage of total shown in parentheses). *Type II de novo contigs*: the total number of *de novo* contigs longer than 49 base pairs, assembled from complexmers that were significantly more frequent in the target genome, given the standardized ratio tests for each pair-wise comparison (see Methods). The mean and maximum size of these contigs is also shown. The codes for samples represent the first letter of the genus and species names, with location number and library type if necessary (see Table 1).

Given the evidence that the reads are not evenly and randomly distributed across the entire genome, as shown in the previous section, a large fraction of these Type I complexmers are probably false positives. Preliminary modeling indicates that ≥20 × coverage is required to obtain reliable results from Type I complexmers. While they are prone to false positives, using an extremely high threshold for an acceptable frequency can still provide a short list of potential markers. We tested the false positive error rate by comparing results for each simulated data sets for a single known genome and found that error only involved low frequency complexmers (results not shown). A more reasonable but still substantial fraction of the Type I complexmers have frequencies above an extremely high threshold (Table 2). Even at this high level of stringency, *Ficus tinctoria* remains remarkably distinct, particularly given the fact that the other three fig samples are PET samples with large amounts of data available.

Type II complexmers

The fraction of complexmers shared between pairs of samples closely followed macro-evolutionary patterns, although there were apparent effects of reproductive behavior on overall genomic divergence. At the population level within a single species, the percent overlap among the ramin individuals ranged from 59% to 99%. The lowest values involved comparisons between SET samples and are probably due largely to poor overlap in the data because of low overall coverage of the genome. On the other hand, the percent overlap between the PET sample (Gb3p) and the other samples was closely correlated with geographic distance, with the peninsular individuals sharing between 93% and 99% of their complexmers while the Sumatran sample shared 89%.

At the family level, ancient and relictual *Trigonobalanus diochangensis* (Manos *et al.* 2001) shared only 6–31% overlap with species of *Lithocarpus*. Here, the SET samples exhibited the highest values of overlap while the PET samples are the most distant. Within the genus *Lithocarpus*, pairwise overlap ranged from 31% to 80%. The lowest value of similarity, between *L. hancei* and *L. xylocarpus*, involved two sympatric but morphologically distinct species that dominate the forests of Ailoshan Mt., Yunnan, China. This value was unusual as the other congeneric similarity values in the stone oaks ranged between 60% and 70%. The other comparisons with greater similarity were all between a SET and PET sample, suggesting a systemic bias in sequence distribution so that same regions of the genome were disproportionately sequenced across samples. Within the figs, the percent overlap among species was substantially lower than within the stone oaks, ranging from 14% to 30%. The percent overlap values of the

pair-wise comparisons were mostly below 20%. Within the genus, the patterns did agree with the general relatedness of the species (Ronsted *et al.* 2005), with *Ficus altissima* and *F. microcarpa* (both in the same subfamily) sharing 27% overlap while *F. fistulosa* and *F. tinctoria* (in closely related subfamilies) shared 30% and comparisons between these two sets of closely related species being quite low (~15%), comparable to the similarity between *Lithocarpus* and its distant outgroup, *T. doichangensis*.

The pair-wise correlation of Type II complexmer frequency also provided an effective method to highlight potential copy number similarities and differences (Fig. 3). At the population level in *Gonystylus bancanus*, the correlation between the Johor PET and Johor SET samples was extremely high and included almost all of the complexmers observed in the SET sample, >2.5 million complexmers (Table 1, Fig. 3a). The bias introduced by the six fold difference in the amount of available sequence data was also apparent (the PET sample on the *y* axis, the SET sample on the *x* axis). At the congeneric level, both the stone oaks (Fig. 3c) and the figs (Fig. 3b) exhibited similar patterns: a broader spread along the 1:1 line and distinct groups of complexmers that were very high frequency in one sample but very low in the other. The correlation between the two stone oak species (*L. xylocarpus* and *L. hancei*) included 6.3 million complexmers while correlation between the two fig species only involved 2.1 million complexmers. At the family level, between *L. xylocarpus* and *T. doichangensis*, the correlation was much weaker with a broad general spread between the two samples and less than 2 million complexmers (Fig. 3d).

In the comparison of the standardized frequency ratios, these general patterns were quite clear (Fig. 4). Between individuals of the same species (Fig. 4a), very few of the Type II complexmers have a strongly skewed frequency ratio, although the Johor SET sample (on the right side of the distribution) does have a surprisingly large number of Type II complexmers with a ratio greater than 10. In general, this ratio test was an effective way to discover potential copy number variants. The targeted *de novo* assembly of both Type II complexmers generated numerous potential informative sequences that represent copy number variants (Table 2). These contigs were generally short, with a

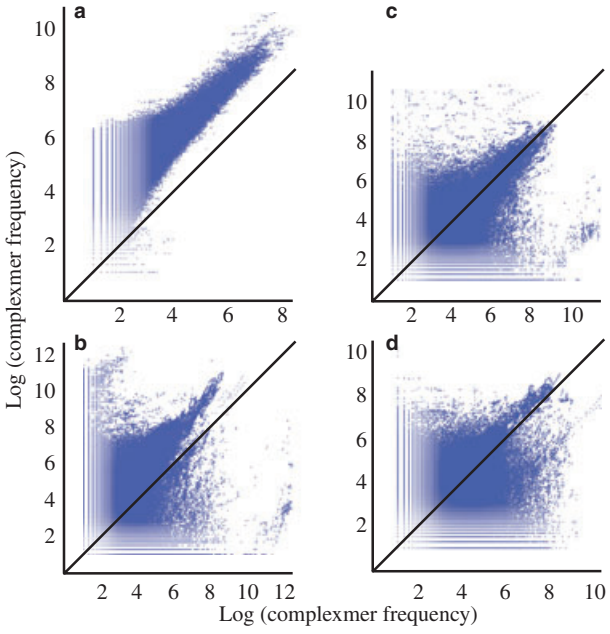

Fig. 3 Correlation of Type II complexmer frequencies between samples representing different macro-evolutionary distances. Values are log-normal transformed and the 1:1 correlation is shown by the solid line. Values for the first member of each pair are shown on the *x*-axis. (a) sympatric individuals within a species: *Gonystylus bancanus* - (Johor) SET vs. (Johor) PET (2 574 024 shared); (b) fig congenerics: *Ficus microcarpa* vs. *F. tinctoria*, both PET (2 106 547 shared); (c) stone oak congenerics: *Lithocarpus hancei* vs. *L. xylocarpus*, both PET (6 016 546 shared); (d) Fagaceae: *L. xylocarpus* vs. *Trigonobalanus doichangiensis*, both PET (1 706 376 shared).

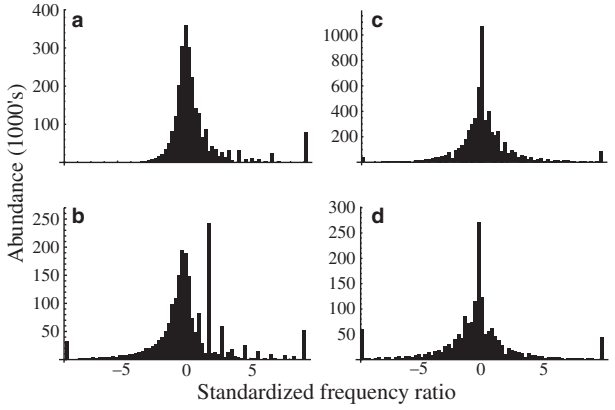

Fig. 4 Distribution of (standardized ratio-1) of Type II complexmer frequencies between samples representing different macro-evolutionary distances. Data was standardized using the median value for each sample. Zero value indicates no difference between samples in the standardized values. Ratios were calculated with the larger value as the numerator to make the distribution symmetrical and all values were greater than one. If the value for the second member of the pair was greater than the first, then the ratio was multiplied by -1. Complexmers that were more frequent in the first member of each pair therefore have positive values. (a) sympatric individuals within a species: *Gonystylus bancanus* - (Johor) SET vs. (Johor) PET; (b) fig congenerics: *Ficus microcarpa* vs. *F. tinctoria*, both PET; (c) stone oak congenerics: *Lithocarpus hancei* vs. *L. xylocarpus*, both PET; (d) Fagaceae: *L. xylocarpus* vs. *Trigonobalanus doichangiensis*, both PET.

mean length ranging from 54 to 74 base pairs, and the longest assemblies ranged from 96 to 461 base pairs.

Type III complexmers

Type III complexmers can be quite powerful in discovering genetic elements that are associated with particular phenotypes and geographic regions. The Type III complexmers were immediately passed into the *de novo* assembly, generating a large number of contigs (Table 3). The most frequent contigs were those which grouped all of the samples within a study group together. The top three groups were all of the figs together, all of the ramin individuals, and the Fagaceae (minus one species). The fourth rank group consisted of all of the Fagaceae. Among the lower frequency contigs were some very interesting groups, particularly for the Fagaceae. One group (Lhanc, Lx, Td) with 90 contigs brought together the Indochinese samples while another (Lt, Lx) with 71 contigs brought together two samples with convergent fruit types, despite the fact that they are found on different continents and one sample (*L. turbinatus*) is represented by SET data. The power of these contigs must be tested in downstream genotyping experiments but our approach generated a large number of candidate sequences.

Table 3 Number of high-quality *de novo* contigs, assembled from Type III complexmers discovered for each group

# of Contigs	Samples in group
662	Fa,Ff,Fm,Ft
231	Cc,Lhanc,Lt,Lx,Td
139	Gb1,Gb2,Gb3s,Gb3p,Gb4
138	Cc,Lhav,Lhanc,Lt,Lx,Td
96	Lhanc,Lt,Lx,Td
91	Gb2,Gb3s,Gb3p,Gb4
90	Lhanc,Lx,Td
78	Cc,Lhanc,Lt,Lx
78	Gb2,Gb3s,Gb3p
71	Ff,Fm,Ft
71	Lt,Lx
64	Fa,Ff
63	Fa,Ff,Ft
60	Lhav,Lhanc,Lt,Lx
60	Cc,Lhanc,Lx,Td
51	Fa,Fm,Ft
51	Ff,Fm
50	Lhanc,Lt
40	Gb1,Gb2,Gb3s,Gb3p
40	Cc,Lhanc,Lx
39	Gb1,Gb3p
36	Cc,Lhav,Lhanc,Lt,Lx
26	Lhanc,Lt,Lx
26	Cc,Lx
21	Gb1,Gb2,Gb3p
21	Lhanc,Td
21	Gb3s,Gb3p

Discussion

For ecologists and evolutionary biologists to quickly and fully incorporate next-gen sequencing technology into their research, a clear distinction needs to be made between 'genome projects' and 'genomic diversity projects'. This distinction is particularly important for tropical biologists, where the most interesting questions involve diverse suites of closely related species, where reproductively isolation is often incomplete. For many purposes, the completion and verification of a physical map of the genome would be prohibitive and, for many purposes, is not necessary. The most important objective is to characterize and identify the genomic differences among exemplar taxa that provide a phenotypic and geographic framework for compelling ecological or evolutionary questions. Here, we demonstrate the power and potential of an approach that discovers relevant DNA sequence variation present in whole genomic short-read data prior to assembly, among genomes at any level of phylogenetic relatedness, by directly comparing the presence and abundance of short complex DNA sequence fragments.

This novel assembly free approach provides several advantages over conventional assembly based approaches, particularly for previously unstudied organisms. Firstly, our approach uses all of the high quality DNA sequence data that passes our simple criteria for sequence complexity and a threshold for complexmer abundance. Unpublished analyses of the SRS data in this study indicate that *de novo* assembly algorithms, such as SOAPdenovo (Li 2009) and ABySS (Langmead *et al.* 2009), incorporate substantially less than half of the available data into high-quality contigs and the results are highly sensitive to *k*-mer length. Secondly, our approach can simultaneously be used to discover differences and identities among genomes, given any subset of phenotypes or geographic origins, at any phylogenetic level of comparison. Thirdly, the overall quality of individual sequencing reactions can be directly compared among lanes, reactions, and samples, without a reference sequence, based upon the dominance diversity curve of complexmer abundance. Finally, direct discovery of relevant genomic differences allows a much smaller and highly focused protocol for the *de novo* assembly of the polymorphisms and their immediate flanking regions. Currently, *de novo* assembly algorithms are optimized to piece together gigabytes of data generated from large entire genomes. To perform this phenomenal task effectively, they must make fairly

conservative decisions about contig extension, allelic diversity, and alternative splicing of the sequence (Warren *et al.* 2007; Simpson *et al.* 2009). We feel that a completely exhaustive search for a short list of informative local assemblies would be possible, using a slightly refined algorithmic approach. This exhaustive approach would map all possible DNA sequence paths flanking the informative complexmers identified in our analysis.

Ultimately, our approach can identify any potential genetic variant within the complexmer window, including single nucleotide polymorphisms, copy number variants, inversions, and insertion-deletion events. While we used a 25 base pair window for this initial analysis, further optimization studies are necessary. Like the choice of *k*-mer length in any alignment or *de novo* algorithm, a balance among specificity and sensitivity must be explored for every analysis. This exploratory step will also be particularly important for comparative studies of genomes at different phylogenetic levels. Preliminary results indicate that 25 base pairs may be too long, providing too much specificity particularly when overall genome coverage is incomplete and generating a large fraction of false positives.

Population-level to family-level applications (and beyond?)

At the population level in an endangered species of tropical timber, we discovered numerous potential markers, including thousands of single nucleotide polymorphisms and hundreds of copy number variants. These markers could be used as DNA fingerprinting tools to regulate the international market for this CITES species. While overall genomic similarities among individuals was high (generally ~96% of complexmers), these results sharply differ from a *de novo* assembly approach that suggested <<0.001% divergence among individuals, based on 10 000 contigs totalling ~1.5 Mb of DNA sequence (C. H. Cannon, unpublished). In this issue, Whittall *et al.* (2009) report very low levels of divergence in the chloroplast genome of the Torrey pine, using de novo assembly techniques, which recovers ~90% of the total genome. It would be interesting to test our approach on these data to find out if the more variable regions are hiding in the remaining 10% of the unassembled genome. This difference in result highlights the fact that our approach uses all of the available data and the bias in *de novo* assembly towards conserved and non-variable regions of the genome.

We also observed that genomic similarity is substantially greater among species of stone oak (*Lithocarpus*) than among fig species (Table 2). The percent overlap among congenerics in the Fagaceae generally ranged between 60% and 70%, even between *Lithocarpus* species sampled on different continents with substantially different fruit morphologies while the fig species almost never shared more than 20% overlap of their complexmers. These substantial differences are probably due to differences in reproductive strategy and gene flow pattern between these two groups. Trees in the Fagaceae are known to hybridize (Manos *et al.* 1999) and previous results from chloroplast DNA sequence revealed abundant shared polymorphisms among species, particularly between sympatric ones (Cannon & Manos 2003). The porous nature of these genomes probably results in broader taxonomic sharing of genomic elements. The figs, on the other hand, have complex co-evolutionary relationships with their wasp pollinators (Janzen 1979; Weiblen 2002). These co-evolutionary interactions (Ronsted *et al.* 2005) probably result in rapid divergence of genomic diversity, although the evidence for the specificity of these co-evolutionary relationships and specificity of genetic markers are not universal (Jackson *et al.* 2008). Two of the fig species are dioecious but no obvious effect of sexual system was apparent. In relation to the original ecological and evolutionary questions outlined in the Methods for the stone oaks, we have identified a long list of genetic elements correlated with fruit type and continental distribution (see Table 3).

Additionally, the analysis of complexmer abundance provides descriptors of overall genomic diversity. For example, the rice genome appears to be substantially more complex, given an equal amount of DNA sequence data, than the poplar genome, despite the relative similarity in genome size between these two species (Fig. 2b). While the two genomes have roughly the same number of total unique complexmers (Fig. 2a), the frequency of these complexmers in rice is several natural-log orders of magnitude greater in abundance than in poplar. Interpreting the differences among the curves for the actual SRS data sets is complicated by obvious differences in the quality and quantity of the sequencing reaction, factors which did not vary in the simulated study. While little shape variation is present among the curves for the simulated studies, the curves for two fig species differ dramatically. The data for *Ficus altissima* has a few complexmers which are extremely abundant but this drops off rapidly so that the 25 000 ranked complexmer is significantly rare. This curve suggests a strong bias in the distribution of the reads towards a smaller fraction of the overall genome. The data for *Ficus tinctoria*, on the other hand, exhibits a curve similar in overall shape to the simulated data curves, only being highly elevated because of the greater overall amount of data available (Table 1). Almost equivalent amounts of data are available for two of the Fagaceae samples as for *F. tinctoria* but the

total numbers of complexmers and the dominance–diversity curves are substantially lower in the Fagaceae. This result may indicate that genome size is considerably larger for the Fagaceae samples or a larger fraction of the genome is non-complex.

The percent overlap in the frequent complexmers among the samples from the three different study groups and the three known genomes, using this approach, are almost negligible, averaging less than 1%. While these comparisons between distantly related species may not have immediate implications for ecological or evolutionary studies, the almost complete divergence in complexmer presence at this deep macroevolutionary level indicates the power and specificity of only using the complex portion of the genome, where virtually no homoplasy is found. In other words, as species diverge, the complex portion of their genomes become increasingly different with very little convergent evolution, which avoids the issue of long branch attraction observed in many studies using conserved genes and partial DNA sequences from a small number of loci (Delsuc *et al.* 2005). These types of markers could quite useful for phylogenetic reconstruction as well, although further development and a deeper understanding of the nature of homology is required.

Next-gen sequencing and tropical biology

Prior to the major breakthroughs in DNA sequencing technology, genomic level studies were limited to a very small subset of model organisms (Ellegren 2008; Holt & Jones 2008). Despite the relatively cheap costs involved with SRS technology, given the massive amounts of data generated, relatively few projects addressing fundamental questions of tropical biodiversity, including genomic diversity and differentiation among suites of closely related species, have been attempted. Tropical communities are the most diverse (Myers *et al.* 2000), most threatened (Jepson *et al.* 2001; Sodhi *et al.* 2004; Butler & Laurance 2008), and among the most complex terrestrial ecosystems on Earth. Many of the questions facing tropical biologists will require the kind of deep insight and understanding provided by genomic scale studies. Given the pace of change, both in terms of land-use (Foley *et al.* 2005), climate (Keller 2009), and threat of species extinction (Vitousek *et al.* 1997), we strongly feel that a major shift is necessary. We suggest that 'genomic diversity' projects focus on 'model groups' of species, possessing important phenotypic and geographic variation in relation to environmental gradients, land-use history, and biogeographic patterns. Exemplar species, representing the overall range of variation in phenotype and geography can be chosen to establish the overall framework. Comparative studies among these exemplar species then reveal the important dimensions of genomic variation associated with these phenotypic traits and geographic locations. As technological and cost barriers continue to fall, we should anticipate a radically different approach to genomic biology.

Our assembly free comparative genomic approach avoids the prior assembly and curation of hundreds of millions of base pairs, in complicated scaffolds of tens of thousand of contigs but instead can directly discover the differences among genomes relevant to the ecological or evolutionary question. The assembly of a previously unknown genome is analogous to reconstructing a library that has been shredded into small fragments of paper. In assembly based analyses, where a final physical map is the prime objective, the library must be completed first, even down to the position of the books on the shelves, before the books can be read. Our assembly free comparative approach, on the other hand, first pinpoints the differences between different but similar libraries to identify specific passages in the vast number of volumes which will provide the most interesting reading. Limiting the bioinformatic challenge of *de novo* assembly to these relatively few number of passages will allow a more exhaustive algorithmic search through all possible variants of the targeted region.

The successful *de novo* assembly of the informative complexmers, which distinguished both individual genomes and groups of related genomes, illustrates the effectiveness of our approach. While direct screening of the complexmers is possible, a simple and traditional approach would be to develop PCR primers for Sanger sequencing of informative fragments. Within these study groups, we identified hundreds of informative fragments which have high quality alignments against the original SRS data for the genomes being distinguished. Most of these fragments contain SNPs or even large pieces of unique sequence, not present in the other genomes. At the population level within ramin, hundreds of high quality contigs grouping different subsets of individuals from different geographic locations were identified. These informative differences could be explored further on high-throughput platforms. Populations of individuals could be screened directly, on various downstream DNA fingerprinting platforms, like microarrays or bead arrays, as breeds of dogs (Barsh 2007) are currently identified. In this volume, Buggs *et al.* (2009) observed a very high discovery rate for valid SNPs in a pair of closely related sunflower species using such a genotyping platform. While the distribution of genetic variation in wild populations will inherently be more complicated, these informative complexmers would create a highly

sensitive and multi-locus DNA barcode. This DNA barcode would be particularly powerful as it would be framed in relation to the ecological and evolutionary contrasts among the focal taxa. Additionally, it would provide a realistic approach to understanding the overall pattern of genomic diversification among suites of closely related species.

Here, we describe a single approach to directly compare previously unstudied genomes, from the population to the family level, without a closely related reference genome or *de novo* assembly of the SRS data. Our approach provides objective measures of the total amount and evenness of complex DNA sequence diversity in the data, providing assessments of data quality and sequencing error rates. Pair-wise comparisons of complexmer frequency between genomes immediately highlights genetic elements which are present in one genome and not the other and points towards regions that have copy number variation. Genetic elements shared among groups of genomes can be used in a targeted *de novo* assembly approach to identify those contigs that most powerfully distinguish the group, in relation to other genomes. Our approach avoids many disadvantages of techniques that require prior assembly, as it uses all of the available data and creates an avenue leading straight to the most informative genomic differences, given the specific comparison. We feel that relatively shallow 'survey' sequencing of numerous species within ecological and economically important groups and subsequent comparative analyses, such as the one described here, will lead to a revolutionary new understanding of how genomes diversify, how species diverge, what functional elements play important ecological and phenotypic roles, and ultimately, how this knowledge can be translated into applied management and monitoring tools for the conservation of tropical biodiversity. While not explored here, our approach would be equally or if not more effective for transcriptome, partial genome, or cytoplasmic genome studies. In those types of studies, complete coverage of the target sequences given a single reaction lane would greatly decrease the false positive error rate and increase the overall sensitivity of pair-wise comparisons.

Acknowledgements

Funding for this research was provided by grants from the National Basic Research Program of China (973 Program, Grant No. 2007CB411600), Chinese Academy of Science's Frontiers in Innovative Research, National Geographic Society's Conservation Trust, the UNDP's Global Environmental Facility Program (MAL/99/G31), and Texas Tech University's Research Enhancement Fund. We thank the Forestry Departments of Johor, Pahang, Sabah, and Terengganu (Malaysia) and the Forestry Department of Indonesia for permission to collect samples. We kindly thank the Diamond Raya Timber Co. (Indonesia) for assistance with the collection of ramin samples in Sumatra, Indonesia. We kindly thank W. Ratnam (National University of Malaysia) and F. Nadia for assistance in sample collection in peninsular Malaysia. Zhou Z-K provided the sample of *Trigonobalanus doichangensis*. Illumina sequencing was performed at Canada's Michael Smith Genome Science Centre at the British Columbia Cancer Agency. Zhao Y-J, R. Warren, S. Jackman, and B. Langmead provided valuable assistance with data analysis and management.

Conflicts of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Amaratunga D, Cabrera J (2004) *Exploration and Analysis of DNA Microarray and Protein Array Data*. John Wiley & Sons, Hoboken, NJ.

Anshari G, Kershaw P, van der Kaars S (2001) A Late Pleistocene and Holocene pollen and charcoal record from peat swamp forest, Lake Sentarum Wildlife Reserve, West Kalimantan, Indonesia. *Palaeogeography Palaeoclimatology Palaeoecology*, **171**, 213–228.

Barsh GS (2007) How the dog got its spots. *Nature Genetics*, **39**, 1304–1306.

Bookjans G, Stummann BM, Henningsen KW (1984) Preparation of chloroplast DNA from pea plastids isolated in a medium of high ionic strength. *Analytical Biochemistry*, **141**, 244–247.

Buggs R.J., Chamala S, Wu W, Gao L, May GD, Schnable PS, Soltis DE, Soltis PS, Barbazuk WB (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology*, **19** (Suppl. 1), 132–146.

Butler RA, Laurance WF (2008) New strategies for conserving tropical forests. *Trends in Ecology & Evolution*, **23**, 469–472.

Cannon CH, Leighton M (2004) Tree species distributions across five habitats in a Bornean rain forest. *Journal of Vegetation Science*, **15**, 257–266.

Cannon CH, Manos PS (2000) The Bornean *Lithocarpus* Bl. section *Synaedrys* (Lindl.) Barnett (Fagaceae): its circumscription and description of a new species. *Botanical Journal of the Linnean Society*, **133**, 343–357.

Cannon CH, Manos PS (2001) Combining and comparing continuous morphometric descriptors with a molecular phylogeny: the case of fruit evolution in the Bornean *Lithocarpus* (Fagaceae). *Systematic Biology*, **50**, 1–21.

Cannon CH, Manos PS (2003) Phylogeography of the Southeast Asian stone oaks (*Lithocarpus*). *Journal of Biogeography*, **30**, 211–226.

Cannon CH, Kua CS, Lobenhofer EK, Hurban P (2006) Capturing genomic signatures of DNA sequence variation using a standard anonymous microarray platform. *Nucleic Acids Research*, **34**. Art. No. e121.

Cannon CH, Morley RJ, Bush ABG (2009) The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. In: *Proceedings of The National Academy of Sciences of The United States of America*, **106**, 11188–11193.

Corlett RT (2007) What's so special about Asian tropical forests? *Current Science*, **93**, 1551–1557.

Cronn R, Liston A, Parks M *et al.* (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*, **36**.

Deguilloux MF, Pemonge MH, Petit RJ (2002) Novel perspectives in wood certification and forensics: dry wood as a source of DNA. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **269**, 1039–1046.

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, **6**, 361–375.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, **36**, e105.

Doyle JJ, Doyle JL (1987) Rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemistry Bulletin*, **19**, 11–15.

Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, **17**, 1629–1631.

Foley JA, DeFries R, Asner GP *et al.* (2005) Global consequences of land use. *Science*, **309**, 570–574.

Harismendy O, Ng PC, Strausberg RL *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**, R32.

Harrison RD (2003) Fig wasp dispersal and the stability of a keystone plant resource in Borneo. *Proceedings of the Royal Society Of London Series B-Biological Sciences*, **270**, S76–S79.

Harrison RD (2005) Figs and the diversity of tropical rainforests. *BioScience*, **55**, 1053–1064.

Harrison RD, Yamamura N (2003) A few more hypotheses for the evolution of dioecy in figs (Ficus, Moraceae). *Oikos*, **100**, 628–635.

Herre EA, Jander KC, Machado CA (2008) Evolutionary ecology of figs and their associates: recent progress and outstanding puzzles. *Annual Review of Ecology Evolution and Systematics*, **39**, 439–458.

Holt RA, Jones SJM (2008) The new paradigm of flow cell sequencing. *Genome Research*, **18**, 839–846.

Hua Z (2008) The tropical flora of southern yunnan, china, and its biogeographic affinities. *Annals Of The Missouri Botanical Garden*, **95**, 661–680.

Huang CJ, Zhang YT, Bartholomew B (2000) Fagaceae. In: *Flora of China: Cycadaceae through Fagaceae* (eds Wu ZY, Raven PH). pp. 314–400. Science Press, Missouri Botanical Garden, Beijing, St Louis.

Jackson AP, Machado CA, Robbins N, Herre EA (2008) Multi-locus phylogenetic analysis of neotropical figs does not support co-speciation with the pollinators: the importance of systematic scale in fig/wasp cophylogenetic studies. *Symbiosis*, **45**, 57–72.

Janzen DH (1979) How to be a fig. *Annual Review of Ecology and Systematics*, **10**, 13–51.

Jepson P, Jarvie JK, MacKinnon K, Monk KA (2001) The end for Indonesia's lowland forests? *Science*, **292**, 859–861.

Keller CF (2009) Global warming: a review of this mostly settled issue. *Stochastic Environmental Research and Risk Assessment*, **23**, 643–676.

Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Li RQ (2009) *Short Oligonucleotide Analysis Package: SOAPdenovo 1.03*. Beijing Genomics Institute, Beijing.

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, **18**, 1851–1858.

Lisitsyn N, Lisitsyn N, Wigler M (1993) Cloning the Differences between 2 Complex Genomes. *Science*, **259**, 946–951.

Machado CA, Robbins N, Gilbert MTP, Herre EA (2005) Critical review of host specificity and its coevolutionary implications in the fig/fig-wasp mutualism. *Proceedings of The National Academy of Sciences of The United States of America*, **102**, 6558–6565.

Manos PS, Doyle JJ, Nixon KC (1999) Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Molecular Phylogenetics and Evolution*, **12**, 333–349.

Manos PS, Zhou ZK, Cannon CH (2001) Systematics of Fagaceae: Phylogenetic tests of reproductive trait evolution. *International Journal of Plant Sciences*, **162**, 1361–1379.

Morley RJ (2000) *Origin and Evolution of Tropical Rain Forests*. John Wiley & Sons, Ltd, New York.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Nielsen LR, Kjaer ED (2008) Tracing timber from forest to consumer with DNA markers (ed. Danish Ministry of the Environment FaNA). http://www.skovognatur.dk/udgivelser.

Petit RJ, Hu FS, Dick CW (2008) Forests of the past: a window to future changes. *Science*, **320**, 1450–1452.

Phillippy AM, Schatz MC, Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology*, **9**, R55.

Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. *Trends In Genetics*, **24**, 142–149.

Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, **27**, 847.

Reinhardt JA, Baltrus DA, Nishimura MT *et al.* (2009) De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Research*, **19**, 294–305.

Rieley JO, Page SE (1995) *Biodiversity and Sustainability of Tropical Peatlands*. Samara Publishing Limited, Cardigan.

Ronsted N, Weiblen GD, Cook JM *et al.* (2005) 60 million years of co-divergence in the fig-wasp symbiosis. *Proceedings of the Royal Society B-Biological Sciences*, **272**, 2593–2599.

Simpson JT, Wong K, Jackman SD *et al.* (2009) ABySS: A parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Smulders MJM, Van Westende WPC, Diway B *et al.* (2008) Development of microsatellite markers in *Gonystylus bancanus* (Ramin) useful for tracing and tracking of wood of this protected species. *Molecular Ecology Resources*, **8**, 168–171.

Sodhi NS, Koh LP, Brook BW, Ng PKL (2004) Southeast Asian biodiversity: an impending disaster. *Trends In Ecology & Evolution*, **19**, 654–660.

Soepadmo E (1972) Fagaceae. In: *Flora Malesiana: Series I—Spermatophytes* (ed. van Steenis CGGJ). Noordhoff International Publishing, Leyden, the Netherlands. 265–403.

Stevens PF (2001 onwards) Angiosperm Phylogeny Website. Version 9, June 2008 [and more or less continuously updated since]. http://www.mobot.org/MOBOT/research/APweb/.

Stromberg MP, Marth GT (2007) *MOSAIK: a reference-guided assembler for next-generation sequence data*. Boston, MA.

van Bers NEM, van Oers K, Kerstens HHD, Dibbits BW, Crooijmans RPMA, Visser ME, Groenen MAM (2010) Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Molecular Ecology*, **19** (Suppl. 1), 89–99.

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM (1997) Human domination of Earth's ecosystems. *Science*, **277**, 494–499.

Warren RL, Sutton GG, Jones SJM, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23**, 500–501.

Weiblen GD (2000) Phylogenetic relationships of functionally dioecious Ficus (Moraceae) based on ribosomal DNA sequences and morphology. *American Journal of Botany*, **87**, 1342–1357.

Weiblen GD (2002) How to be a fig wasp. *Annual Review of Entomology*, **47**, 299–330.

Weiblen GD (2004) Correlated evolution in fig pollination. *Systematic Biology*, **53**, 128–139.

Whittall J, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R (2009) Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology*, **19** (Suppl. 1), 100–114.

Wyn LT, Soehartono T, Keong CH (2004) *Framing the picture: an assessment of ramin trade in Indonesia, Malaysia, and Singapore*. p. 129. TRAFFIC Southeast Asia, Kuala Lumpur.

Xiao ZS, Zhang ZB, Wang YS (2005) Effects of seed size on dispersal distance in five rodent-dispersed fagaceous species. *Acta Oecologica-International Journal Of Ecology*, **28**, 221–229.

Xiao ZS, Wang YS, Harris M, Zhang ZB (2006) Spatial and temporal variation of seed predation and removal of sympatric large-seeded species in relation to innate seed traits in a subtropical forest, Southwest China. *Forest Ecology and Management*, **222**, 46–54.

Young SS, Herwitz SR (1995) Floristic diversity and co-occurrences in a subtropical broad-leaved forest and two contrasting regrowth stands in central-west Yunnan Province, China. *Vegetation*, **119**, 1–14.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Zheng Z, Li QY (2000) Vegetation, climate, and sea level in the past 55,000 years, Hanjiang Delta, southeastern China. *Quaternary Research*, **53**, 330–340.

Zhou ZK, Gilbert MG (2003) Moraceae. In: *Flora of China* (eds Huang H, Raven P). pp. 21–73. Science Press, Missouri Botanical Garden, Beijing, St Louis.

Charles Cannon studies the ecology, evolution, and conservation of tropical Asian forests using a variety of approaches. Currently, he is committed to extending genomic techniques to the study of tropical biodiversity. Chai-Shian Kua has a background in human cancer genetics and is now focused on the application of cutting-edge technologies in non-model systems. Zhang Di helped develop the software for this analysis and has a growing interest in bioinformatics and genomics. John Harting is working on general multivariate evolutionary theory, using stochastic models based upon Price's theorem. He continues to be involved in conservation research in Indonesia.
